# Learning for Microblogs with Distant Supervision: Political Forecasting with Twitter

**Micol Marchetti-Bowick**
Microsoft Corporation
475 Brannan Street
San Francisco, CA 94122
micolmb@microsoft.com

**Nathanael Chambers**
Department of Computer Science
United States Naval Academy
Annapolis, MD 21409
nchamber@usna.edu

## Abstract

Microblogging websites such as Twitter offer a wealth of insight into a population's current mood. Automated approaches to identify general sentiment toward a particular topic often perform two steps: *Topic Identification* and *Sentiment Analysis*. Topic Identification first identifies tweets that are relevant to a desired topic (e.g., a politician or event), and Sentiment Analysis extracts each tweet's attitude toward the topic. Many techniques for Topic Identification simply involve selecting tweets using a keyword search. Here, we present an approach that instead uses distant supervision to train a classifier on the tweets returned by the search. We show that distant supervision leads to improved performance in the Topic Identification task as well in the downstream Sentiment Analysis stage. We then use a system that incorporates distant supervision into both stages to analyze the sentiment toward President Obama expressed in a dataset of tweets. Our results better correlate with Gallup's Presidential Job Approval polls than previous work. Finally, we discover a surprising baseline that outperforms previous work without a Topic Identification stage.

## 1 Introduction

Social networks and blogs contain a wealth of data about how the general public views products, campaigns, events, and people. Automated algorithms can use this data to provide instant feedback on what people are saying about a topic. Two challenges in building such algorithms are (1) identifying topic-relevant posts, and (2) identifying the attitude of each post toward the topic. This paper studies distant supervision (Mintz et al., 2009) as a solution to both challenges. We apply our approach to the problem of predicting Presidential Job Approval polls from Twitter data, and we present results that improve on previous work in this area. We also present a novel baseline that performs remarkably well without using topic identification.

Topic identification is the task of identifying text that discusses a topic of interest. Most previous work on microblogs uses simple keyword searches to find topic-relevant tweets on the assumption that short tweets do not need more sophisticated processing. For instance, searches for the name "Obama" have been assumed to return a representative set of tweets about the U.S. President (O'Connor et al., 2010). One of the main contributions of this paper is to show that keyword search can lead to noisy results, and that the same keywords can instead be used in a distantly supervised framework to yield improved performance.

Distant supervision uses noisy signals in text as positive labels to train classifiers. For instance, the token "Obama" can be used to identify a series of tweets that discuss U.S. President Barack Obama. Although searching for token matches can return false positives, using the resulting tweets as positive training examples provides supervision from a distance. This paper experiments with several diverse sets of keywords to train distantly supervised classifiers for topic identification. We evaluate each classifier on a hand-labeled dataset of political and apolitical tweets, and demonstrate an improvement in F1 score over simple keyword search (.39 to .90 in the best case). We also make available the first labeled dataset for topic identification in politics to encourage future work.

Sentiment analysis encompasses a broad field of research, but most microblog work focuses on two moods: positive and negative sentiment.

*Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*

1. REPORT DATE
**APR 2012**

2. REPORT TYPE

3. DATES COVERED
**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE
**Learning for Microblogs with Distant Supervision: Political Forecasting with Twitter**

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
**Microsoft Corporation,475 Brannan Street,San Francisco,CA,94122**

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

**Microblogging websites such as Twitter offer a wealth of insight into a population?s current mood. Automated approaches to identify general sentiment toward a particular topic often perform two steps: Topic Identification and Sentiment Analysis. Topic Identification first identifies tweets that are relevant to a desired topic (e.g., a politician or event), and Sentiment Analysis extracts each tweet?s attitude toward the topic. Many techniques for Topic Identification simply involve selecting tweets using a keyword search. Here we present an approach that instead uses distant supervision to train a classifier on the tweets returned by the search. We show that distant supervision leads to improved performance in the Topic Identification task as well in the downstream Sentiment Analysis stage. We then use a system that incorporates distant supervision into both stages to analyze the sentiment toward President Obama expressed in a dataset of tweets. Our results better correlate with Gallup?s Presidential Job Approval polls than previous work. Finally, we discover a surprising baseline that outperforms previous work without a Topic Identification stage.**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | **Same as Report (SAR)** | **10** | |

Algorithms to identify these moods range from matching words in a sentiment lexicon to training classifiers with a hand-labeled corpus. Since labeling corpora is expensive, recent work on Twitter uses emoticons (i.e., ASCII smiley faces such as *:-(* and *:-)*) as noisy labels in tweets for distant supervision (Pak and Paroubek, 2010; Davidov et al., 2010; Kouloumpis et al., 2011). This paper presents new analysis of the downstream effects of topic identification on sentiment classifiers and their application to political forecasting.

Interest in measuring the political mood of a country has recently grown (O'Connor et al., 2010; Tumasjan et al., 2010; Gonzalez-Bailon et al., 2010; Carvalho et al., 2011; Tan et al., 2011). Here we compare our sentiment results to Presidential Job Approval polls and show that the sentiment scores produced by our system are positively correlated with both the *Approval* and *Disapproval* job ratings.

In this paper we present a method for coupling two distantly supervised algorithms for topic identification and sentiment classification on Twitter. In Section 4, we describe our approach to topic identification and present a new annotated corpus of political tweets for future study. In Section 5, we apply distant supervision to sentiment analysis. Finally, Section 6 discusses our system's performance on modeling Presidential Job Approval ratings from Twitter data.

## 2 Previous Work

The past several years have seen sentiment analysis grow into a diverse research area. The idea of sentiment applied to microblogging domains is relatively new, but there are numerous recent publications on the subject. Since this paper focuses on the microblog setting, we concentrate on these contributions here.

The most straightforward approach to sentiment analysis is using a sentiment *lexicon* to label tweets based on how many sentiment words appear. This approach tends to be used by applications that measure the general mood of a population. O'Connor et al. (2010) use a ratio of positive and negative word counts on Twitter, Kramer (2010) counts lexicon words on Facebook, and Thelwall (2011) uses the publicly available SentiStrength algorithm to make weighted counts of keywords based on predefined polarity strengths.

In contrast to lexicons, many approaches instead focus on ways to train supervised classifiers. However, labeled data is expensive to create, and examples of Twitter classifiers trained on hand-labeled data are few (Jiang et al., 2011). Instead, distant supervision has grown in popularity. These algorithms use emoticons to serve as semantic indicators for sentiment. For instance, a sad face (e.g., *:-(*) serves as a noisy label for a negative mood. Read (2005) was the first to suggest emoticons for UseNet data, followed by Go et al. (Go et al., 2009) on Twitter, and many others since (Bifet and Frank, 2010; Pak and Paroubek, 2010; Davidov et al., 2010; Kouloumpis et al., 2011). Hashtags (e.g., *#cool* and *#happy*) have also been used as noisy sentiment labels (Davidov et al., 2010; Kouloumpis et al., 2011). Finally, multiple models can be blended into a single classifier (Barbosa and Feng, 2010). Here, we adopt the emoticon algorithm for sentiment analysis, and evaluate it on a specific domain (politics).

Topic identification in Twitter has received much less attention than sentiment analysis. The majority of approaches simply select a single keyword (e.g., "Obama") to represent their topic (e.g., "US President") and retrieve all tweets that contain the word (O'Connor et al., 2010; Tumasjan et al., 2010; Tan et al., 2011). The underlying assumption is that the keyword is precise, and due to the vast number of tweets, the search will return a large enough dataset to measure sentiment toward that topic. In this work, we instead use a distantly supervised system similar in spirit to those recently applied to sentiment analysis.

Finally, we evaluate the approaches presented in this paper on the domain of politics. Tumasjan et al. (2010) showed that the results of a recent German election could be predicted through frequency counts with remarkable accuracy. Most similar to this paper is that of O'Connor et al. (2010), in which tweets relating to President Obama are retrieved with a keyword search and a sentiment lexicon is used to measure overall approval. This extracted approval ratio is then compared to Gallup's Presidential Job Approval polling data. We directly compare their results with various distantly supervised approaches.

## 3 Datasets

The experiments in this paper use seven months of tweets from Twitter (www.twitter.com) collected between June 1, 2009 and December 31, 2009. The corpus contains over 476 million tweets labeled with usernames and timestamps, collected through Twitter's 'spritzer' API without keyword filtering. Tweets are aligned with polling data in Section 6 using their timestamps.

The full system is evaluated against the publicly available daily Presidential Job Approval polling data from Gallup[1]. Every day, Gallup asks 1,500 adults in the United States about whether they approve or disapprove of "the job President Obama is doing as president." The results are compiled into two trend lines for *Approval* and *Disapproval* ratings, as shown in Figure 1. We compare our positive and negative sentiment scores against these two trends.

## 4 Topic Identification

This section addresses the task of **Topic Identification** in the context of microblogs. While the general field of topic identification is broad, its use on microblogs has been somewhat limited. Previous work on the political domain simply uses keywords to identify topic-specific tweets (e.g., O'Connor et al. (2010) use "Obama" to find presidential tweets). This section shows that *distant supervision* can use the same keywords to build a classifier that is much more robust to noise than approaches that use pure keyword search.

### 4.1 Distant Supervision

Distant supervision uses noisy signals to identify positive examples of a topic in the face of unlabeled data. As described in Section 2, recent sentiment analysis work has applied distant supervision using emoticons as the signals. The approach extracts tweets with ASCII smiley faces (e.g., *:)* and *;)*) and builds classifiers trained on these positive examples. We apply distant supervision to topic identification and evaluate its effectiveness on this subtask.

As with sentiment analysis, we need to collect positive and negative examples of tweets about the target topic. Instead of emoticons, we extract positive tweets containing one or more predefined keywords. Negative tweets are randomly chosen from the corpus. Examples of positive and negative tweets that can be used to train a classifier based on the keyword "Obama" are given here:

> **positive:** *LOL, obama made a bears reference in green bay. uh oh.*
>
> **negative:** *New blog up! It regards the new iPhone 3G S: <URL>*

| ID | Type | Keywords |
|---|---|---|
| PC-1 | Obama | *obama* |
| PC-2 | General | *republican*, *democrat*, *senate*, *congress*, *government* |
| PC-3 | Topic | *health care*, *economy*, *tax cuts*, *tea party*, *bailout*, *sotomayor* |
| PC-4 | Politician | *obama*, *biden*, *mccain*, *reed*, *pelosi*, *clinton*, *palin* |
| PC-5 | Ideology | *liberal*, *conservative*, *progressive*, *socialist*, *capitalist* |

Table 1: The keywords used to select positive training sets for each political classifier (a subset of all PC-3 and PC-5 keywords are shown to conserve space).

We then use these automatically extracted datasets to train a multinomial Naive Bayes classifier. Before feature collection, the text is normalized as follows: (a) all links to photos (twitpics) are replaced with a single generic token, (b) all non-twitpic URLs are replaced with a token, (c) all user references (e.g., *@MyFriendBob*) are collapsed, (d) all numbers are collapsed to INT, (e) tokens containing the same letter twice or more in a row are condensed to a two-letter string (e.g. the word *ahhhhh* becomes *ahh*), (f) lowercase the text and insert spaces between words and punctuation. The text of each tweet is then tokenized, and the tokens are used to collect unigram and bigram features. All features that occur fewer than 10 times in the training corpus are ignored.

Finally, after training a classifier on this dataset, every tweet in the corpus is classified as either positive (i.e., relevant to the topic) or negative (i.e., irrelevant). The positive tweets are then sent to the second sentiment analysis stage.

### 4.2 Keyword Selection

Keywords are the input to our proposed distantly supervised system, and of course, the input to previous work that relies on keyword search. We evaluate classifiers based on different keywords to measure the effects of keyword selection.

O'Connor et al. (2010) used the keywords "Obama" and "McCain", and Tumasjan et al. (2010) simply extracted tweets containing Germany's political party names. Both approaches extracted matching tweets, considered them rele-

[1] http://gallup.com/poll/113980/gallup-daily-obama-job-approval.aspx

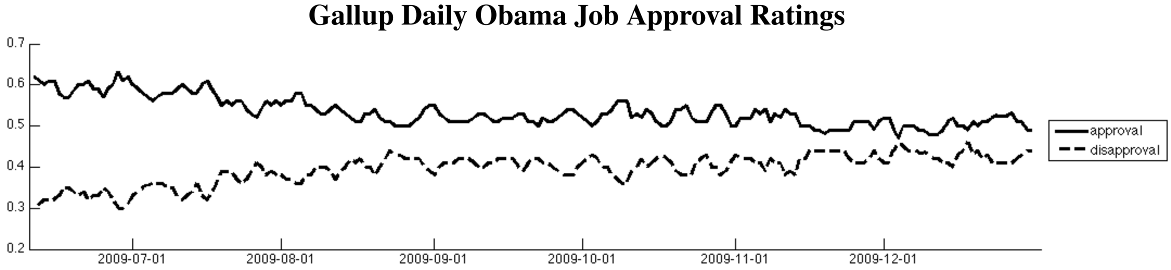

Figure 1: Gallup presidential job Approval and Disapproval ratings measured between June and Dec 2009.

vant (correctly, in many cases), and applied sentiment analysis. However, different keywords may result in very different extractions. We instead attempted to build a generic "political" topic classifier. To do this, we experimented with the five different sets of keywords shown in Table 1. For each set, we extracted all tweets matching one or more keywords, and created a balanced positive/negative training set by then selecting negative examples randomly from non-matching tweets. A couple examples of ideology (PC-5) extractions are shown here:

> *You often hear of deontologist libertarians and utilitarian* ***liberals*** *but are there any Aristotelian socialists?*
>
> *<url> - Then, slather on a* ***liberal*** *amount of plaster, sand down smooth, and paint however you want. I hope this helps!*

The second tweet is an example of the noisy nature of keyword extraction. Most extractions are accurate, but different keywords retrieve very different sets of tweets. Examples for the political topics (PC-3) are shown here:

> *RT @PoliticalMath: hope the president's* ***health care*** *predictions <url> are better than his stimulus predictions <url>*
>
> *@adamjschmidt You mean we could have chosen* ***health care*** *for every man woman and child in America or the Iraq war?*

Each keyword set builds a classifier using the approach described in Section 4.1.

### 4.3 Labeled Datasets

In order to evaluate distant supervision against keyword search, we created two new labeled datasets of political and apolitical tweets.

The **Political Dataset** is an amalgamation of all four keyword extractions (PC-1 is a subset of PC-4) listed in Table 1. It consists of 2,000 tweets randomly chosen from the keyword searches of PC-2, PC-3, PC-4, and PC-5 with 500 tweets from each. This combined dataset enables an evaluation of how well each classifier can identify tweets from other classifiers. The **General Dataset** contains 2,000 random tweets from the entire corpus. This dataset allows us to evaluate how well classifiers identify political tweets in the wild.

This paper's authors initially annotated the same 200 tweets in the General Dataset to compute inter-annotator agreement. The Kappa was 0.66, which is typically considered *good* agreement. Most disagreements occurred over tweets about money and the economy. We then split the remaining portions of the two datasets between the two annotators. The Political Dataset contains 1,691 political and 309 apolitical tweets, and the General Dataset contains 28 political tweets and 1,978 apolitical tweets. These two datasets of 2000 tweets each are publicly available for future evaluation and comparison to this work[2].

### 4.4 Experiments

Our first experiment addresses the question of keyword variance. We measure performance on the Political Dataset, a combination of all of our proposed political keywords. Each keyword set contributed to 25% of the dataset, so the evaluation measures the extent to which a classifier identifies other keyword tweets. We classified the 2000 tweets with the five *distantly supervised* classifiers and the one "Obama" *keyword* extractor from O'Connor et al. (2010).

Results are shown on the left side of Figure 2. Precision and recall calculate correct identification of the political label. The five distantly supervised approaches perform similarly, and show remarkable robustness despite their different training sets. In contrast, the keyword extractor only

[2] http://www.usna.edu/cs/nchamber/data/twitter

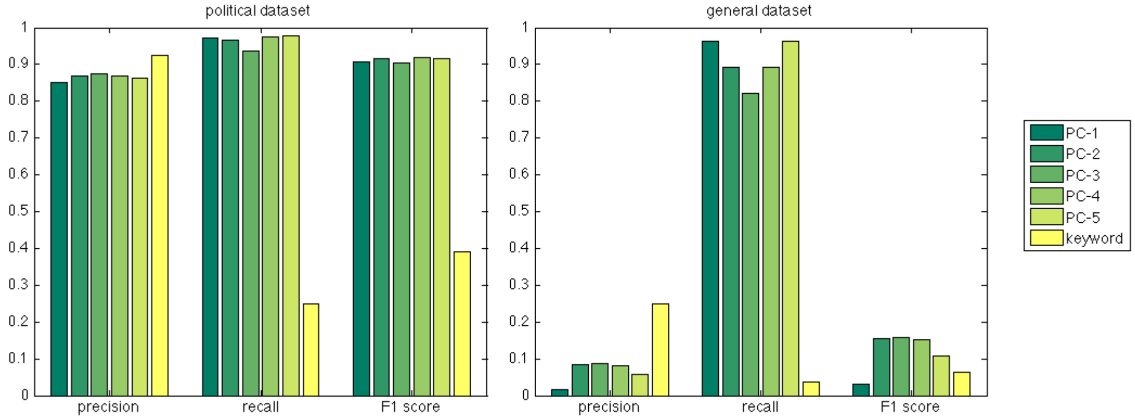

Figure 2: Five distantly supervised classifiers and the Obama keyword classifier. Left panel: the Political Dataset of political tweets. Right panel: the General Dataset representative of Twitter as a whole.

captures about a quarter of the political tweets. PC-1 is the distantly supervised analog to the Obama keyword extractor, and we see that distant supervision increases its F1 score dramatically from 0.39 to 0.90.

The second evaluation addresses the question of classifier performance on Twitter as a whole, not just on a political dataset. We evaluate on the General Dataset just as on the Political Dataset. Results are shown on the right side of Figure 2. Most tweets posted to Twitter are not about politics, so the apolitical label dominates this more representative dataset. Again, the five distant supervision classifiers have similar results. The Obama keyword search has the highest precision, but drastically sacrifices recall. Four of the five classifiers outperform keyword search in F1 score.

### 4.5 Discussion

The Political Dataset results show that distant supervision adds robustness to a keyword search. The distantly supervised "Obama" classifier (PC-1) improved the basic "Obama" keyword search by 0.51 absolute F1 points. Furthermore, distant supervision doesn't require additional human input, but simply adds a trained classifier. Two example tweets that an Obama keyword search misses but that its distantly supervised analog captures are shown here:

> *Why does Congress get to opt out of the Obummercare and we can't. A company gets fined if they don't comply. Kiss freedom goodbye.*
>
> *I agree with the lady from california, I am sixty six years old and for the first time in my life I am ashamed of our government.*

These results also illustrate that distant supervision allows for flexibility in construction of the classifier. Different keywords show little change in classifier performance.

The General Dataset experiment evaluates classifier performance in the wild. The keyword approach again scores below those trained on noisy labels. It classifies most tweets as apolitical and thus achieves very low recall for tweets that are actually about politics. On the other hand, distant supervision creates classifiers that over-extract political tweets. This is a result of using balanced datasets in training; such effects can be mitigated by changing the training balance. Even so, four of the five distantly trained classifiers score higher than the raw keyword approach. The only underperformer was PC-1, which suggests that when building a classifier for a relatively broad topic like politics, a variety of keywords is important.

The next section takes the output from our classifiers (i.e., our topic-relevant tweets) and evaluates a fully automated sentiment analysis algorithm against real-world polling data.

## 5 Targeted Sentiment Analysis

The previous section evaluated algorithms that extract topic-relevant tweets. We now evaluate methods to distill the overall sentiment that they express. This section compares two common approaches to sentiment analysis.

We first replicated the technique used in O'Connor et al. (2010), in which a lexicon of positive and negative sentiment words called OpinionFinder (Wilson and Hoffmann, 2005) is used to evaluate the sentiment of each tweet (others have used similar lexicons (Kramer, 2010; Thelwall et al., 2010)). We evaluate our full distantly supervised approach to theirs. We also experimented with SentiStrength, a lexicon-based program built to identify sentiment in online comments of the social media website, MySpace. Though MySpace is close in genre to Twitter, we did not observe a performance gain. All reported results thus use OpinionFinder to facilitate a more accurate comparison with previous work.

Second, we built a distantly supervised system using tweets containing emoticons as done in previous work (Read, 2005; Go et al., 2009; Bifet and Frank, 2010; Pak and Paroubek, 2010; Davidov et al., 2010; Kouloumpis et al., 2011). Although distant supervision has previously been shown to outperform sentiment lexicons, these evaluations do not consider the extra topic identification step.

### 5.1 Sentiment Lexicon

The OpinionFinder lexicon is a list of 2,304 positive and 4,151 negative sentiment terms (Wilson and Hoffmann, 2005). We ignore neutral words in the lexicon and we do not differentiate between *weak* and *strong* sentiment words. A tweet is labeled positive if it contains any positive terms, and negative if it contains any negative terms. A tweet can be marked as both positive and negative, and if a tweet contains words in neither category, it is marked neutral. This procedure is the same as used by O'Connor et al. (2010). The sentiment scores $S_{pos}$ and $S_{neg}$ for a given set of $N$ tweets are calculated as follows:

$$S_{pos} = \frac{\sum_x 1\{x_{label} = positive\}}{N} \quad (1)$$

$$S_{pos} = \frac{\sum_x 1\{x_{label} = negative\}}{N} \quad (2)$$

where $1\{x_{label} = positive\}$ is 1 if the tweet $x$ is labeled positive, and $N$ is the number of tweets in the corpus. For the sake of comparison, we also calculate a **sentiment ratio** as done in O'Connor et al. (2010):

$$S_{ratio} = \frac{\sum_x 1\{x_{label} = positive\}}{\sum_x 1\{x_{label} = negative\}} \quad (3)$$

### 5.2 Distant Supervision

To build a trained classifier, we automatically generated a positive training set by searching for tweets that contain at least one positive emoticon and no negative emoticons. We generated a negative training set using an analogous process. The emoticon symbols used for positive sentiment were *:) =) :-) :] =] :-] :} :o) :D =D :-D :P =P :-P C:*. Negative emoticons were *:( =( :-( :[ =[ :-[ :{ :-c :c} D: D= :S :/ =/ :-/ :'( :_(*. Using this data, we train a multinomial Naive Bayes classifier using the same method used for the political classifiers described in Section 4.1. This classifier is then used to label topic-specific tweets as expressing positive or negative sentiment. Finally, the three overall sentiment scores $S_{pos}$, $S_{neg}$, and $S_{ratio}$ are calculated from the results.

## 6 Predicting Approval Polls

This section uses the two-stage Targeted Sentiment Analysis system described above in a real-world setting. We analyze the sentiment of Twitter users toward U.S. President Barack Obama. This allows us to both evaluate distant supervision against previous work on the topic, and demonstrate a practical application of the approach.

### 6.1 Experiment Setup

The following experiments combine both topic identification and sentiment analysis. The previous sections described six topic identification approaches, and two sentiment analysis approaches. We evaluate all combinations of these systems, and compare their final sentiment scores for each day in the nearly seven-month period over which our dataset spans.

Gallup's Daily Job Approval reports two numbers: Approval and Disapproval. We calculate individual sentiment scores $S_{pos}$ and $S_{neg}$ for each day, and compare the two sets of trends using Pearson's correlation coefficient. O'Connor et al. do not explicitly evaluate these two, but instead use the ratio $S_{ratio}$. We also calculate this daily ratio from Gallup for comparison purposes by dividing the Approval by the Disapproval.

### 6.2 Results and Discussion

The first set of results uses the lexicon-based classifier for sentiment analysis and compares the different topic identification approaches. The first table in Table 2 reports Pearson's correlation coefficient with Gallup's Approval and Disapproval ratings. Regardless of the Topic classifier, all

**Sentiment Lexicon**

| Topic Classifier | Approval | Disapproval |
|---|---|---|
| keyword | -0.22 | 0.42 |
| PC-1 | -0.65 | 0.71 |
| PC-2 | -0.61 | 0.71 |
| PC-3 | -0.51 | 0.65 |
| PC-4 | -0.49 | 0.60 |
| PC-5 | -0.65 | 0.74 |

**Distantly Supervised Sentiment**

| Topic Classifier | Approval | Disapproval |
|---|---|---|
| keyword | 0.27 | 0.38 |
| PC-1 | **0.71** | **0.73** |
| PC-2 | 0.33 | 0.46 |
| PC-3 | 0.05 | 0.31 |
| PC-4 | 0.08 | 0.26 |
| PC-5 | 0.54 | 0.62 |

Table 2: Correlation between Gallup polling data and the extracted sentiment with a lexicon (trends shown in Figure 3) and distant supervision (Figure 4).

**Sentiment Lexicon**

| keyword | PC-1 | PC-2 | PC-3 | PC-4 | PC-5 |
|---|---|---|---|---|---|
| .22 | .63 | .46 | .33 | .27 | .61 |

**Distantly Supervised Sentiment**

| keyword | PC-1 | PC-2 | PC-3 | PC-4 | PC-5 |
|---|---|---|---|---|---|
| .40 | .64 | .46 | .30 | .28 | .60 |

Table 3: Correlation between Gallup Approval / Disapproval ratio and extracted sentiment ratio scores.

systems inversely correlate with Presidential Approval. However, they correlate well with Disapproval. Figure 3 graphically shows the trend lines for the keyword and the distantly supervised system PC-1. The visualization illustrates how the keyword-based approach is highly influenced by day-by-day changes, whereas PC-1 displays a much smoother trend.

The second set of results uses distant supervision for sentiment analysis and again varies the topic identification approach. The second table in Table 2 gives the correlation numbers and Figure 4 shows the keyword and PC-1 trend lines.The results are widely better than when a lexicon is used for sentiment analysis. Approval is no longer inversely correlated, and two of the distantly supervised systems strongly correlate (PC-1, PC-5).

The best performing system (PC-1) used distant supervision for both topic identification and sentiment analysis. Pearson's correlation coefficient for this approach is 0.71 with Approval and 0.73 with Disapproval.

Finally, we compute the ratio $S_{ratio}$ between the positive and negative sentiment scores (Equation 3) to compare to O'Connor et al. (2010). Table 3 shows the results. The distantly supervised topic identification algorithms show little change between a sentiment lexicon or a classifier. However, O'Connor et al.'s keyword approach improves when used with a distantly supervised sentiment classifier (.22 to .40). Merging Approval and Disapproval into one ratio appears to mask the sentiment lexicon's poor correlation with Approval. The ratio may not be an ideal evaluation metric for this reason. Real-world interest in Presidential Approval ratings desire separate Approval and Disapproval scores, as Gallup reports. Our results (Table 2) show that distant supervision avoids a negative correlation with Approval, but the ratio hides this important advantage.

One reason the ratio may mask the negative Approval correlation is because tweets are often classified as both positive and negative by a lexicon (Section 5.1). This could explain the behavior seen in Figure 3 in which both the positive and negative sentiment scores rise over time. However, further experimentation did not rectify this pattern. We revised $S_{pos}$ and $S_{neg}$ to make binary decisions for a lexicon: a tweet is labeled positive if it strictly contains more positive words than negative (and vice versa). Correlation showed little change. Approval was still negatively correlated, Disapproval positive (although less so in both), and the ratio scores actually dropped further. The sentiment ratio continued to hide the poor Approval performance by a lexicon.

### 6.3 New Baseline: Topic-Neutral Sentiment

Distant supervision for sentiment analysis outperforms that with a sentiment lexicon (Table 2). Distant supervision for topic identification further improves the results (PC-1 v. keyword). The best system uses distant supervision in both stages (PC-1 with distantly supervised sentiment), outperforming the purely keyword-based algorithm of O'Connor et al. (2010). However, the question of how important topic identification is has not yet been addressed here or in the literature.

Both O'Connor et al. (2010) and Tumasjan et al. (2010) created joint systems with two topic identification and sentiment analysis stages. But

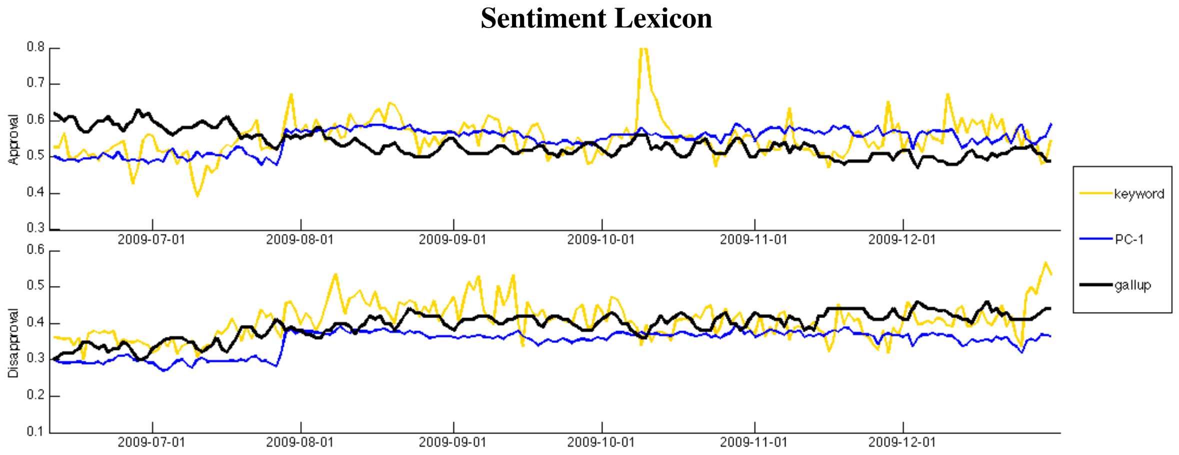

Figure 3: Presidential job approval and disapproval calculated using two different topic identification techniques, and using a sentiment lexicon for sentiment analysis. Gallup polling results are shown in black.

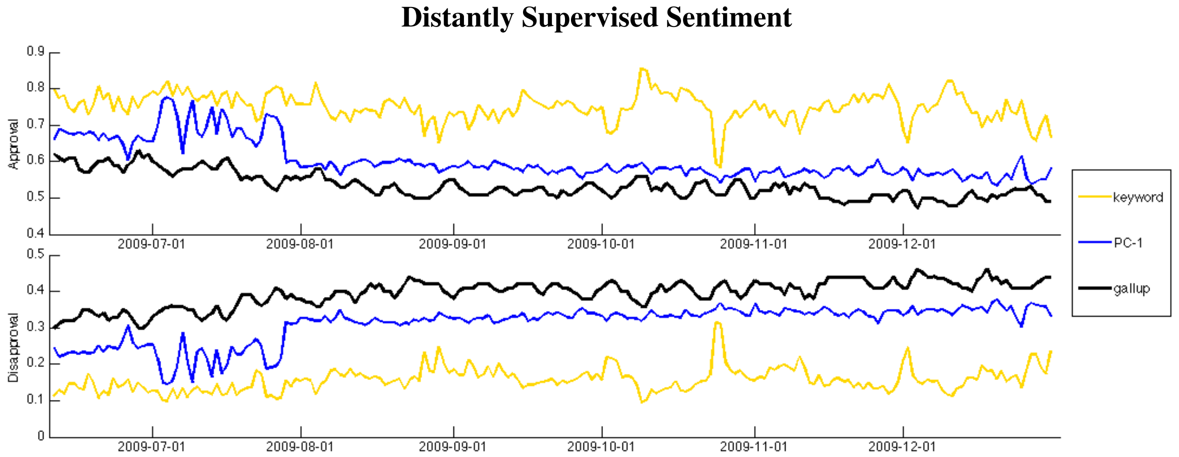

Figure 4: Presidential job approval sentiment scores calculated using two different topic identification techniques, and using the emoticon classifier for sentiment analysis. Gallup polling results are shown in black.

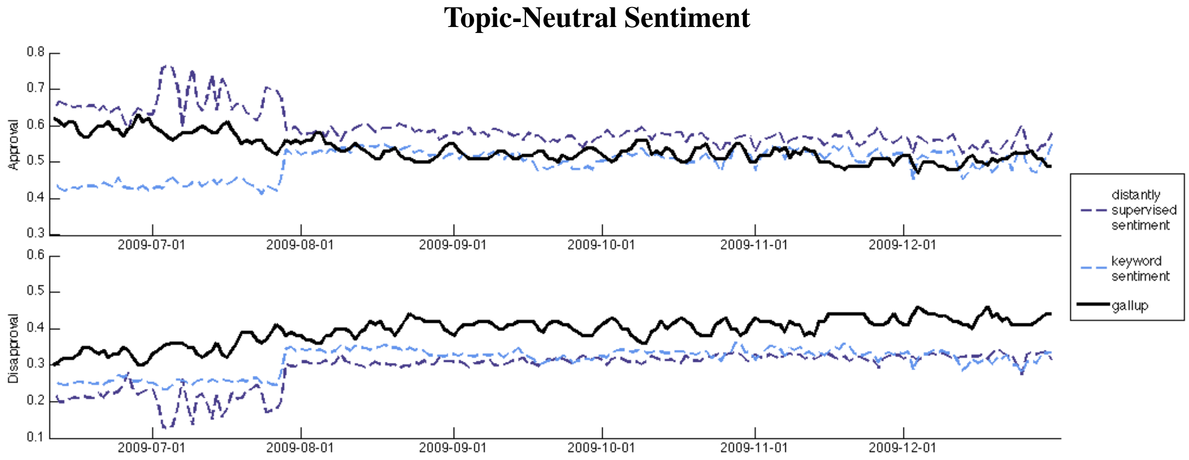

Figure 5: Presidential job approval sentiment scores calculated using the entire twitter corpus, with two different techniques for sentiment analysis. Gallup polling results are shown in black for comparison.

**Topic-Neutral Sentiment**

| Algorithm | Approval | Disapproval |
|---|---|---|
| Distant Sup. | **0.69** | **0.74** |
| Keyword Lexicon | -0.63 | 0.69 |

Table 4: Pearson's correlation coefficient of Sentiment Analysis without Topic Identification.

what if the topic identification step were removed and sentiment analysis instead run on the entire Twitter corpus? To answer this question, we ran the distantly supervised emoticon classifier to classify *all* tweets in the 7 months of Twitter data. For each day, we computed the positive and negative sentiment scores as above. The evaluation is identical, except for the removal of topic identification. Correlation results are shown in Table 4.

This baseline parallels the results seen when topic identification is used: the sentiment lexicon is again inversely correlated with Approval, and distant supervision outperforms the lexicon approach in both ratings. This is not surprising given previous distantly supervised work on sentiment analysis (Go et al., 2009; Davidov et al., 2010; Kouloumpis et al., 2011). However, our distant supervision also performs as well as the best performing *topic-specific* system. The best performing topic classifier, PC-1, correlated with Approval with $r$=0.71 (0.69 here) and Disapproval with $r$=0.73 (0.74 here). Computing overall sentiment on Twitter performs as well as political-specific sentiment. This unintuitive result suggests a new baseline that all topic-based systems should compute.

## 7 Discussion

This paper introduces a new methodology for gleaning topic-specific sentiment information. We highlight four main contributions here.

First, this work is one of the first to evaluate distant supervision for topic identification. All five political classifiers outperformed the lexicon-driven keyword equivalent that has been widely used in the past. Our model achieved .90 F1 compared to the keyword .39 F1 on our political tweet dataset. On twitter as a whole, distant supervision increased F1 by over 100%. The results also suggest that performance is relatively insensitive to the specific choice of seed keywords that are used to select the training set for the political classifier.

Second, the sentiment analysis experiments build upon what has recently been shown in the literature: distant supervision with emoticons is a valuable methodology. We also expand upon prior work by discovering drastic performance differences between positive and negative lexicon words. The OpinionFinder lexicon failed to correlate (inversely) with Gallup's Approval polls, whereas a distantly trained classifier correlated strongly with both Approval and Disapproval (Pearson's .71 and .73). We only tested OpinionFinder and SentiStrength, so it is possible that another lexicon might perform better. However, our results suggest that lexicons vary in their quality across sentiment, and distant supervision may provide more robustness.

Third, our results outperform previous work on Presidential Job Approval prediction (O'Connor et al., 2010). We presented two novel approaches to the domain: a coupled distantly supervised system, and a topic-neutral baseline, both of which outperform previous results. In fact, the baseline surprisingly matches or outperforms the more sophisticated approaches that use topic identification. The baseline correlates .69 with Approval and .74 with Disapproval. This suggests a new baseline that should be used in all topic-specific sentiment applications.

Fourth, we described and made available two new annotated datasets of political tweets to facilitate future work in this area.

Finally, Twitter users are not a representative sample of the U.S. population, yet the high correlation between political sentiment on Twitter and Gallup ratings makes these results all the more intriguing for polling methodologies. Our specific 7-month period of time differs from previous work, and thus we hesitate to draw strong conclusions from our comparisons or to extend implications to non-political domains. Future work should further investigate distant supervision as a tool to assist topic detection in microblogs.

## Acknowledgments

We thank Jure Leskovec for the Twitter data, Brendan O'Connor for open and frank correspondence, and the reviewers for helpful suggestions.

## References

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010).*

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *Lecture Notes in Computer Science*, volume 6332, pages 1–15.

Paula Carvalho, Luis Sarmento, Jorge Teixeira, and Mario J. Silva. 2011. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the Association for Computational Linguistics (ACL-2011)*, pages 564–568.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010).*

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report.

Sandra Gonzalez-Bailon, Rafael E. Banchs, and Andreas Kaltenbrunner. 2010. Emotional reactions and the pulse of public opinion: Measuring the impact of political events on the sentiment of online discussions. Technical report.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the Association for Computational Linguistics (ACL-2011).*

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.*

Adam D. I. Kramer. 2010. An unobtrusive behavioral model of ‘gross national happiness’. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI 2010).*

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL ’09, pages 1003–1011.

Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the AAAI Conference on Weblogs and Social Media.*

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference On Language Resources and Evaluation (LREC).*

Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop (ACL-2005).*

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review.*

J.; Wilson, T.; Wiebe and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing.*